Annotation of vector genomes: The *Aedes aegypti* model

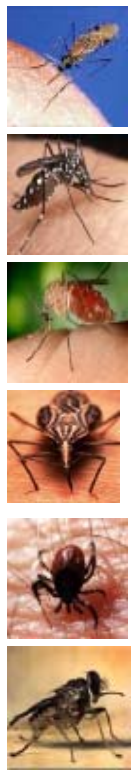
Daniel Lawson, Frank Collins

BRC 6th February 2006

VectorBase

Vectors being sequenced

- *Anopheles gambiae* (PEST, M & S strains)
- *Aedes aegypti*
- *Culex pipiens quinquefasciatus*
- *Rhodnius prolixus* [NHGRI]
- *Ixodes scapularis*
- *Glossina morsitans morsitans* [Sanger/Japan]



VectorBase

Overview

- Generating sequence
- Annotation plan
- Gene prediction
- Dissemination of data (publication/web)☐
- Commitment to curation effort

Generating sequence

- Sequencing undertaken by established sequencing centres

  (e.g. Broad, GSC, JCVI, Sanger,TIGR)
- Initial assembly to be annotated in collaboration with the sequencing centre(s)

Annotation plan

- First-pass gene prediction
  - Focused on protein-coding genes CDS's
- Semi-automated approach
  - This is not manual curation
- Involvement of community where possible
- Timely delivery of gene set

Gene Prediction

- Each group/centre has it's own gene prediction pipeline/protocol.
- Each group produces a 1st pass 'best guess' set of predictions
- These sets are merged into a canonical set
- Which is annotated with protein features
- .. And released to the wider world

Aedes aegypti as an example of VectorBase curation

- *A.aegypti* has a 1.3 Gb genome
- Genomic sequencing was a collaboration between TIGR & Broad
- Assembly 1.0 release in August 2005
- Preliminary gene sets (0.5) December 2005
- Final gene set (1.0) February 2006

Curation efforts

- Broad
  - 10 Mb region for manual annotation
- TIGR
  - Full gene build
- VectorBase
  - Full gene build

Preparing for gene build

- RepeatMask
  - Analyses to identify repeat elements
    - RepeatScout
    - RECON
  - Standard tandem-repeat & low-complexity filtering
- Collate data sets
  - Transcripts (cDNA & EST data)
  - Peptides (Aedes, and taxonomic groupings)
- Train gene predictors

Merge of data sets to 1.0 release

- Simple, hierarchical system
- Reduce to single transcript per locus (simplicity)
- Compare loci across the 2 sets
- Categorize
- Manually investigate some examples
- Deal with each category in a different manner
- Collate each group back to give a ‘minimal’ complete set
- Add alternate isoforms back into the set (transcripts, proteins)
- Add UTR extensions where possible
- QC the data set

Examples of categories

ISOFORM-SAME

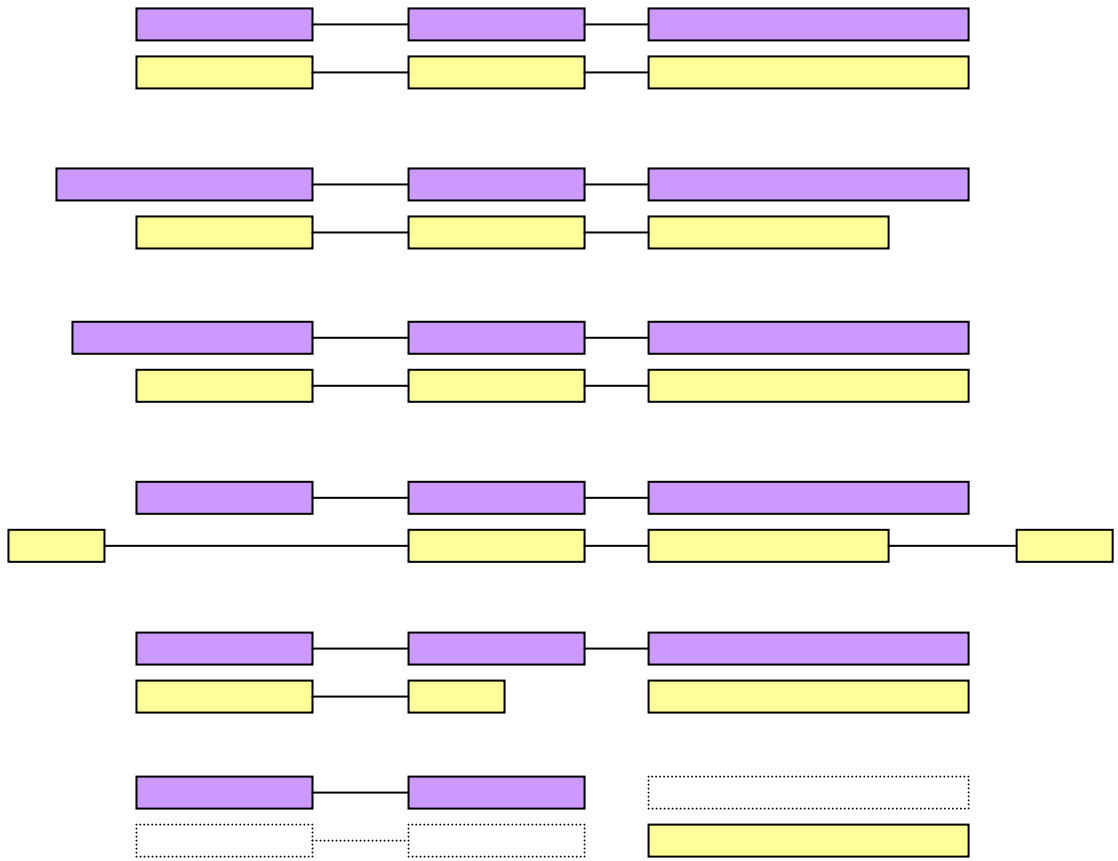
COMPAT-ENCAPS

COMPAT-END_AGREE

ISOFORM-DIFFERENT

SPLIT or MERGE

SINGLETON (NOMAP)



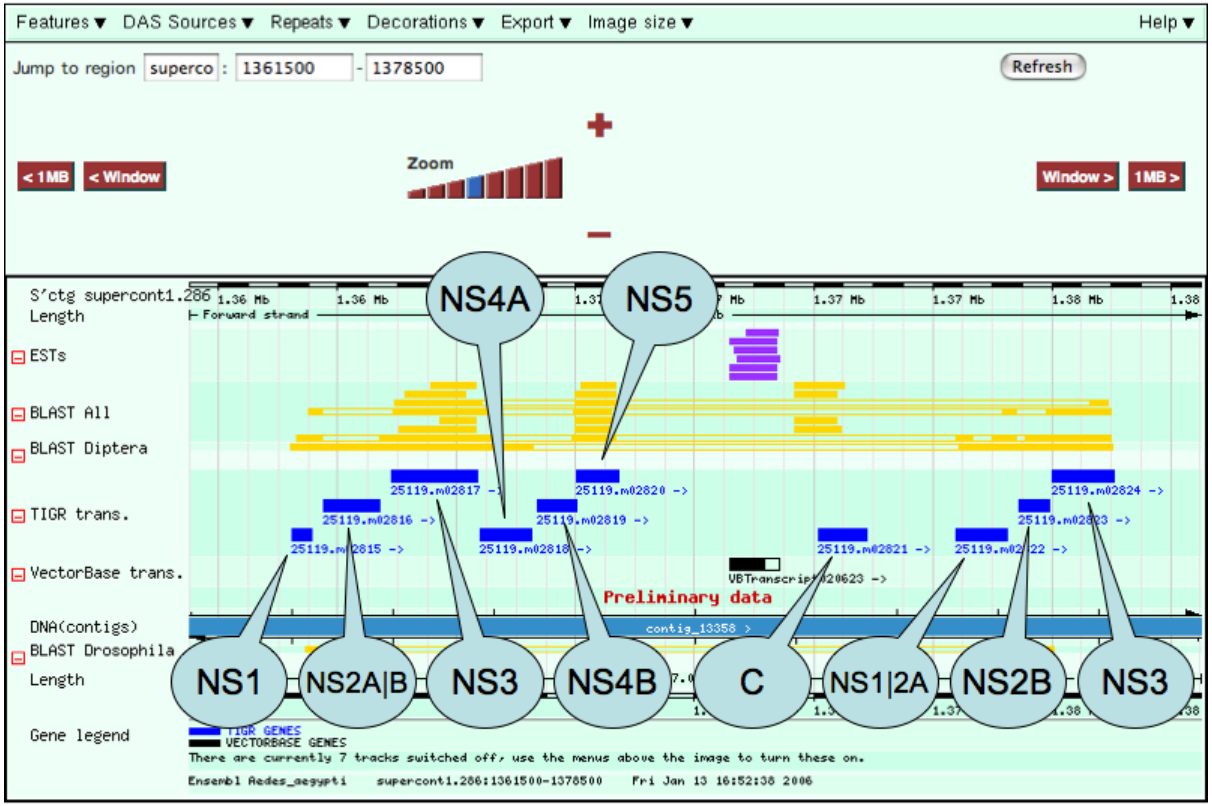
VectorBase

Generation of 0.5 Gene sets

- VectorBase and TIGR produced a CDS prediction set

VectorBase	TIGR
17,776 genes	28,301 genes
21,431 transcripts	29,569 transcripts

Flaviviral integrations in the Aedes genome



- Integration of segments of a flavivaral genome are rare
- ORFs are frame-shifted and appear non-functional

VectorBase

Protein features

- Using the 1.0 gene set
- Look for protein features
  - Domains (InterPro/Panther/TIGRfam)
  - Signal peptides
  - Trans-membrane domains
  - Low-complexity regions
- Give generic description where possible

Data release

- Web-based genome browser
  - VectorBase
  - EnsEMBL
- Flatfiles available from FTP
  - BRC
- Submission to GenBank/EMBL/DDBJ

Curation v Annotation

- VectorBase is committed to ongoing curation of the vector genomes
- Genome maintenance
  - Updates to genome sequence (Yearly)
  - Updates to gene set (6 months)
- Incorporation of other data types
  - Microarray expression data
  - RNAi knockdown data

Acknowledgements

- Sequencers:
  - TIGR
  - Broad Institute
- Annotation:
  - EMBL-EBI
    - Martin Hammond
    - Karyn Megy
    - GeneBuild team
  - TIGR
    - Jennifer Wortman
    - Brian Haas
    - Joshua Orvis
    - Linda Hannick
    - Shelby Bidwell
  - Broad
    - Chinnappa Kodira

- Web:
  - Sanger:
    - James Smith
    - Fiona Cunningham
  - VectorBase (Notre Dame):
    - EO Stinson
    - Rob Bruggner
- BRC:
  - Todd Creasey (TIGR)